A Study of Machine Learning Approaches for Integrated Biomedical Data Analysis

Yi-Tan Chang

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Engineering

Guoqiang Yu, Chairman
Yue Wang
Lamine Mili

May 7th, 2018
Arlington, Virginia

Keywords: Data integration, machine learning, pathway enrichment, pathway prioritization, matrix completion, treatment recommendation.

A Study of Machine Learning Approaches for Integrated Biomedical Data Analysis

Yi-Tan Chang

ABSTRACT

This thesis consists of two projects in which various machine learning approaches and statistical analysis for the integration of biomedical data analysis were explored, developed and tested. Integration of different biomedical data sources allows us to get a better understating of human body from a bigger picture. If we can get a more complete view of the data, we not only get a more complete view of the molecule basis of phenotype, but also possibly can identify abnormality in diseases which were not found when using only one type of biomedical data. The objective of the first project is to find biological pathways which are related to Duechenne Muscular Dystrophy(DMD) and Lamin A/C(LMNA) using the integration of multi-omics data. We proposed a novel method which allows us to integrate proteins, mRNAs and miRNAs to find disease related pathways. The goal of the second project is to develop a personalized recommendation system which recommend cancer treatments to patients. Compared to the traditional way of using only users' rating to impute missing values, we proposed a method to incorporate users' profile to help enhance the accuracy of the prediction.

A Study of Machine Learning Approaches for Integrated Biomedical Data Analysis

Yi-Tan Chang

GENERAL AUDIENCE ABSTRACT

There are two existing major problems in the biomedical field. Previously, researchers only used one data type for analysis. However, one measurement does not fully capture the processes at work and can lead to inaccurate result with low sensitivity and specificity. Moreover, there are too many missing values in the biomedical data. This left us with many questions unanswered and can lead us to draw wrong conclusions from the data. To overcome these problems, we would like to integrate multiple data types which not only better captures the complex biological processes but also leads to a more comprehensive characterization. Moreover, utilizing the correlation among various data structures also help us impute missing values in the biomedical datasets.

For my two research projects, we are interested in integrating multiple biological data to identify disease specific pathways and predict unknown treatment responses for cancer patients. In this thesis, we propose a novel approach for pathways identification using the integration of multi-omics data. Secondly, we also develop a recommendation system which combines different types of patients' medical information for missing treatment responses' prediction. Our goal is that we would find disease related pathways for the first project and enhance missing treatment response's prediction for the second project with the

methods we develop.

The findings of my studies show that it is possible to find pathways related to muscular dystrophies using the integration of multi-omics data. Moreover, we also demonstrate that incorporating patient's genetic profile can improve the prediction accuracy compared to using the treatment responses matrix alone for imputation.

Acknowledgments

I would first like to thank my advisor, Dr. Guoqiang Yu, and Dr. Yue Wang for giving me such amazing research topics as my thesis projects. I have learnt so much in the past two years. Not only have I grown a lot academically, but also as a person. They have taught me the persistence and intuition needed when conducting research. I greatly appreciate them for continuing to support me throughout my time in Virginia Tech.

Secondly, I would also like to thank Dr. Lamine Mili for being in my committee and taking the time to attend my thesis defense.

Thirdly, I would like to thank my supervisor, Dr. Joseph Bender, and Dr. Subha Madhavan for giving me the opportunity to worked at Perthera, Inc. over the summer and continue my research at the company as part of my thesis topic. I appreciate them for always taking the time out of their busy schedule to sit down with me and walk through my research problem.

Lastly, I would like to thank my lab mates at CBIL for their kindness and patience in teaching me many mathematical and statistical concepts I had trouble understanding.

Table of Contents

List of Figures

List of Tables

Chapter 1: Integrated Pathway Prioritization

1.1 Introduction

The muscular dystrophies are inborn errors of metabolism resulting in muscle weakness, wasting, and often lead to an early death. To understand the cause of dystrophy, it is important to identify the biological pathways which are related to the disease. Identifying what genes, proteins and other molecules are involved in a biological pathway can provide clues about what goes wrong in the pathways. Hence, it will help us cure the disease when it strikes and prevent the disease from striking in the future. Traditionally, most biologists only use single omic data for pathways identification. This is mainly due to the lack of computational power in the past to handle large scale biological data and the complex biological relationship between molecules. Although using single omic for finding disease related pathways is cost-effective and the corresponding analytical approaches are well established, we are likely to miss important biological information. For the project, we developed algorithms to enable the multi-omics analysis and integration. The goal is to use the multi-omics data to find disease related pathway. The reason why we want to use multi-omic data instead of single omic data is that integrating multi-omic data into finding pathways provides us a more comprehensive overview of otherwise fragmented information on how molecules truly function in our body. Hence, by doing this, we will get a better understanding of the dysregulation of biomarkers leading to the disease.

The procedure of using multi-omics data to find pathways can be divided into 4 steps and the multi-omics data we want to integrate are proteomic, mRNA and miRNA data. First, we want to identify the proteins and mRNA that are expressed differently in our disease of interest compared

to normal group and the rest of the 6 diseases groups. Since pathways are a series of genes, miRNA names cannot be found in pathways. Hence, in order to utilize the miRNA information, we need to find the proteins and mRMAs which are negatively regulated by miRNAs since the inverse relation between genes and miRNAs is biologically expected. To determine how significant a pathway is based on disease specific proteins and mRNAs as well as proteins and mRNAs which are negatively correlated to miRNAs, hypergeometric distribution is applied for gene enrichment analysis to find significant pathways. Finally, we would like to integrate the overlapped pathways we found for proteins and mRNAs on each of the four platforms using Fishers' method. The combined p-values we obtained from Fishers' method is what we use to determine which pathways are related to the disease under investigation.

1.2 Background

Many studies have been conducted mRNA proteins and miRNAs for pathway analysis. In the past, most pathways' analysis was done on either proteins or mRNAs. This is mainly due to the high level of complexity within the regulatory relationship between these 3 molecules and the lack of technology in the past to process a large amount of data. Using proteins and mRNA to find disease specific pathways, the most common way to do this is through gene enrichment analysis. Gene enrichment analysis is a method to identify genes or proteins that are over-represented in a large group dataset which might be associated with disease phenotype. This method is commonly used by biologists to find differentially expressed genes which are significant enriched in their associated pathways. Only a very few pathways analysis were done on miRNAs or the integration of all three omes. To identify pathways associated to a list of miRNAs, genes targeted by any miRNA of interest needed to be first identified by reference

database or prediction algorithms. Then the significance of the overlap between target genes and pathway genes is measured by an enrichment analysis. Pathways which are associated with miRNAs are the pathways which the enriched genes are targeted by at least one miRNA according to MetaBase, TargetScan or mirTarBase[1]. Moreover, in [2], the author was interested in finding disease risk related pathways regulated by miRNAs. They developed a pathway identification method, called MRPP (miRNA regulated risk pathways by sample-matched miRNA-mRNA profiles). This method incorporated sample-matched miRNA-mRNA expression profiles and pathway structure information. Furthermore, a recent solution was proposed in 2017[3]. It explored two approaches to identify cancer-related miRNAs and investigate relationships between miRNAs and the regulatory networks in cancer. They proposed to use proteins and mRNAs which are both significantly correlated with target miRNAs as inputs for finding cancer related pathways. Secondly, they also presented a different method using SAMBA bi-clustering algorithm [21] and a Bayesian network model for the integration of protein information before the addition of miRNAs to the modules. Even though these methods have considered the regulatory impact of miRNAs in the regulation of the pathway, they failed to integrate both proteins and mRNAs which are also disease specific for analyzing disease related pathways. Even though miRNAs are responsible for the regulation of target genes involved in many biological processes, only considering the proteins or mRNAs which are anti-correlated with their target miRNA neglects the proteins and mRNAs which are not regulated by miRNAs but still behave abnormally among patients. Hence, we proposed methods which have taken the 1) complex regulation of mRNA, proteins and miRNA as well as 2) stringent mathematical models to find disease specific mRNAs and proteins separately with the goal to detect disease specific pathways.

1.3 Methodology

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product, which is usually protein. A gene is considered to be differentially expressed if there is a statistically significant change in expression level between the case and control group. These kinds of changes can be measured through different statistical methods using statistical distributional property of the data. In the project, the case will be a group of people with our disease of interest and the control group is the group of people with no diseases or the combination of 6 other groups of disease which is not in our disease of interest. The reason why we compared the case group with two control groups separately is that we want to compare our disease of interest in a no-disease condition and separately with multiple diseases combined to find out the genes that are being significantly expressed in the disease of our interest. We proposed three methods which allow us to find out these significantly different genes in the group of our disease of interest. These methods will be discussed in detailed in the following sections.

Before applying any of the statistical methods to find out disease specific genes, we first remove the proteins which contain missing values in our samples. Since there are only 5 patients in each of the experimental groups, a low sample numbers resulting from missing values can cause low statistical power and negatively affect the likelihood that a true statistically significant gene being detected. Moreover, we also removed mRNAs with extremely low expression value across all 8 different experimental groups. We considered the extremely low expression value occurs due to the noise produced by machine.

1.3.1 Finding Disease Specific Genes

1.3.1.1 Using OVE-PUG to Find Disease Specific Genes

Identification of markers associated with a particular subpopulation has always been a fundamental challenge for biologists. To find disease specific genes, we adopted a method named one-versus-everyone fold change(OVE-FC) to find genes that are expressed differently only in our diseases of interest. The original goal of OVE-FC is to overcome the problem of using one-versus-rest fold change (OVR-FC) to find significant molecular markers. Using OVR-FC, one would compare the ratio of the expression mean in one specific phenotype to the grand mean of all the rest of the phenotypic groups. Thus, some genes can be mistaken as markers due to the large difference in expression level among unbalanced samples numbers in different phenotypic groups. For example, when calculating the grand mean for "the rest of the phenotypic group", a molecular with much lower expression intensity in just a few phenotypic groups is still likely to have high OVR-FC due to the averaging effect and can be mistakenly selected as markers. To overcome this problem, Yu, et al. [15] proposed to rank markers using OVE-FC (Eq. 2.2), which achieves much better classification performance and also helps detect novel markers in multiple applications. OVE-FC assures selected markers are highly expressed in one phenotypic group relatively to each of the remaining group. Our group proposed a new rigorous multiple group comparison procedure, called OVE-PUG (phenotypic upregulated marker genes) test. It adopts the concept of OVE-FC to find the significant phenotypic upregulated marker genes by computing one-versus-everyone PUG-statistic and estimates its null distribution by weighted permutation scheme.

OVE-FC in here is defined as:

$$OVE - FC(j,k) = \frac{\overline{s_k}(j)}{\max_{l \neq k} \overline{s_l}(j)}$$

for the $\overline{s_k}$ is the geometric mean of expression levels of jth molecule under phenotype k.

Since all three omes of our data are logged, and the corresponding test statistic for OVE-LFC(one-versus-everyone log fold change) standardized by variance is:

$$t_{jk} = \min_{l \neq k} \left\{ \frac{\hat{\mu}_{jk} - \hat{\mu}_{jl}}{\hat{\sigma}_j = \sqrt{\frac{1}{N_k} + \frac{1}{N_l}}} \right\} = \min_{l \neq k} \{t - stat_{k,l}(j)\}, k = 1, \ldots, K$$

where $\hat{\sigma}_j{}^2$is the estimated genewise variance, N_k is the sample number under phenotype k, N_l is the sample number under phenotype l and the arithmetic mean of $log\ \overline{s_{ij}}(j)$ is $\hat{\mu}_{jk} = log\ \overline{s_k}\ (j)$.

$stat_{k,l}(j)$ is t-statistic between phenotype k and l. We call t_{jk} as OVE PUG-statistic associated with phenotype k. The null distribution of the OVE PUG-statistic is the likelihood that the expected expression level of 1 gene being expressed the same in top two or more disease groups.

After the estimation of p-values under each phenotypic group, molecule-wise multiple testing correction needs to be performed. Each molecule is actually tested K times although we only record one p-value associated with highest expressed phenotype. Bonferroni correction could be applied here.

1.3.1.2 Using T-test and Fold-Change to Obtain Differentially Expressed Genes

For this approach, we used unpaired T-test and log fold changes to find out disease specific genes between disease vs. normal and disease vs. all the rest of the diseases' group. For applying for T-test, we Unpaired T-test is a commonly applied statistical approach, which allows us to

compare the means of two unmatched groups, assuming that the case and control groups roughly follow a normal distribution. The test statistic for unpaired T-test is:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \overline{x_1})^2 + \sum_{j=1}^{n_2}\left(x_j - \overline{x_2}\right)^2}{n_1 + n_2 - 2}$$

$\overline{x_1}$ and $\overline{x_2}$ = samples means between the case and control groups

s^2 = sample variance

n_1 and n_2 = sample sizes for case and control group

t = the student t quantile with $n_1 + n_2 - 2$ degrees of freedom

The T-test would return a p-value which represents how likely that two groups are not significantly different taking into account their mean difference, variance and also their sample size. However, since we conducted T-test on case and control groups for 962 genes, we applied multiple testing correction on the p-value obtained from T-test to control the expected proportion of "discoveries" that are false. The chosen multiple testing correction here is false discovery rate(FDR). False discovery rate is the number of false positives over the sum of false positive and true positives. False positives are the genes which are not true markers but are indicated as significant in the test results. Ture positives are the genes which are true markers and also found statistically significant. The Benjamini–Hochberg(B-H) procedure is used to control the FDR under any given threshold α. A brief procedure of B-H procedure is shown as follows:

Multiple null hypothesis $H_1 \ldots H_m$ are conducted on multiple genes and $P_1 \ldots P_m$ are their corresponding p-values to represent the significance of a gene being differentially expressed.

1. Rank each individual p-values in ascending order.
2. Assign ranks to the p-values based on their ascending order. The smallest p-value has rank 1. The rank increases as p-values increase.
3. Calculate each individual p-value's Benjamini-Hochberg critical value, using the formula $\frac{i}{m}Q$, where:

   i = the individual p-value's rank

   m = total number of tests

   Q = False discovery rate

For a given Q, we want to find the largest i such that $P_k < \frac{i}{m}Q$. Then we would reject the null hypothesis for all $H_{(i)}$ for $i = 1, \ldots, k$. Moreover, an adjusted p-value, called q-value, is also calculated for each gene using the Benjamini-Hochberg approach.

$$q_i = p_i * \frac{m}{i}$$

P-values estimate the fraction of null results that are classified as significant, q-values estimate the fraction of significant results that are actually null. For example, p-value of 0.05 implies that 5% of all tests will result in false positives and q-value of 0.05 implies that 5% of significant tests will result in false positives. The q-value for a result is the minimum false positive rate for all results scoring at least as well as the given result. In general, q-values will result in less false positives.

A protein or mRNA which has q value less than some thresholds and fold change greater than some thresholds in both disease vs. norm and disease vs. rest of all other disease groups combined are considered to be significantly up-regulated differentially expressed genes.

1.3.1.3 Using Sigmoid Function

A sigmoid function is a mathematical function having a characteristic "S"-shaped curve. It is commonly used in neural network as an activation function of neurons to output the neuron values between 0 and 1. We modified the original sigmoid function to scaled into a hyperbolic tangent function so the range of the outputs is now between -1 and 1. We want to mapped those genes with extreme log fold change to be on the two ends of the tanh function and those log fold change with smaller values to stay near zero in the tanh graph. Then, we can set thresholds to determine the differentially expressed genes which are considered significant.

$$S(x) = \frac{2}{1 + e^{-\frac{x}{c}}} - 1$$

x = log fold change values of genes

c = scaling parameters to determine how steep the sigmoid function is

The parameter c determines how steep the tanh function is. The smaller c is, the steeper tanh function is. The tanh function will gene to push genes with larger fold change to the 1 or -1. Genes with smaller fold change will tend to stay closer to 0 in the tanh function. We can set different thresholds on the fold change outputs from the tanh function so that genes with fold change greater than the threshold are considered differentially expressed.

1.3.2 Integrate Multiple miRNAs for Each Corresponding Gene

1.3.2.1 Using Target Score from miRDB

miRNA plays a critical role in gene regulation because they acts as a post-transcriptional gene expression regulators and are involved in several important physiological processes such as development, cell differentiation and cell signaling[10] The regulation between miRNAs and genes is very complicated. Several thousand human genes are potential targets for regulation by the several hundred miRNAs encoded in the genome. There are also many genetic confounding factors and the lack of strong biological evidence makes it challenging to accurately predict miRNA targets.

miRDB[12,13] is an online database for predicting miRNA targets in animals. miRNAs. In miRDB, all the predicted targets have target prediction scores ranged between 50 and 100. These scores are reliably computed using miRNA target prediction program based on support vector machines (SVMs) and high-throughput training datasets. Empirical evidences show that a predicted target with prediction score higher than 80 is most likely to be real. However, a gene candidate with score less than 60 is considered to be an unlikely target.

One miRNA may regulate many genes as its targets. We want to obtain the overall log fold-change of each gene taking account into the prediction score of the gene regulated by their multiple corresponding miRNA. We used the weighted average to combine the weighted miRNA log fold change for each gene. The weights here are the target prediction score of miRNA for each gene.

$$Over\ all\ Fold\ Change\ of\ corresponding\ miRNAs\ for\ gene\ i = \frac{\sum_{i=0}^{m} w_i s_i}{\sum_{i=0}^{m} w_i}$$

s = log fold change of corresponding miRNAs on each gene

w = target score of each gene and its corresponding miRNA

m = total number of corresponding miRNA for each gene

Using weighted average to calculate overall fold change of each genes regulated by its corresponding miRNA is a more accurate measurement than using arithmetic mean to calculate the combined log fold change of miRNAs since it takes into account that not all corresponding miRNAs contribute the same regulation possibility for each target gene.

1.3.2.2 Using Correlation Coefficient

We are interested in finding out the miRNA-proteins and miRNA-mRNAs which follow the biologically expected regulation pattern. The expected regulation pattern is that the regulation direction of genes and miRNA is the opposite of each other. More specifically, it is expected that the downregulation of miRNA will lead to the upregulation of genes. However, in some rare cases, the upregulation of miRNA will also lead to the downregulation of genes.

The regulatory relationship between miRNA and gene is complex since one miRNA may regulate many genes as its targets, and at the same time one gene may be targeted by many miRNAs. Rather than solving this multiple-to-multiple relationship. We are more interested in the genes which has inverse regulatory relationship with its corresponding miRNAs. To solve this problem, we take the average of the expression of the corresponding miRNA for each gene. This is because multiple miRNAs regulate one gene. The Pearson's correlation coefficient of each gene and the average of its corresponding miRNA is calculated as follows:

For each gene,

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2) - (n\sum Y^2 - (\sum Y)^2)}}$$

X = gene expression of 5 normal and 5 disease patients

Y = average miRNAs expression of 5 normal and 5 disease patients

To determine whether the correlation between of each and their corresponding miRNAs is statistically significant, we also examine the p-value of the correlation between each pair of miRNA-gene. The p-value of Pearson's correlation coefficient follows the t-distribution and it is calculated as follows:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The p-value is 2 × P(T > t) where T follows a *t* distribution with $n - 2$ degrees of freedom.

r = correlation coefficient

n = number of observations

Since multiple hypothesis are performed on each corresponding miRNA-gene pair, FDR is again applied to control the numbers of false positives in the test results. If a gene has negative correlation and q-value is significant under the chosen threshold with its corresponding miRNA, we consider that this gene is significantly negatively correlated with its corresponding miRNAs.

1.3.2.3 Using Hypergeometric Distribution

The third approach is to use hypergeometric distribution[16] to identify significant genes which are inversely regulated by their corresponding miRNAs. For a given gene, the significance level was calculated as follows:

$$p = 1 - \sum_{k=0}^{m} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

M = number of miRNA negatively regulating this gene among all the miRNAs in the dataset

m = number of corresponding miRNA negatively regulating this gene

N = total number of miRNA in the dataset

n = number of corresponding miRNA regulating this gene

Each gene will be given a p-value representing how likely this gene is being negatively regulated by more than m number of miRNAs. FDR will again be applied to the p-values for each gene to correct the false positive rate for the genes we found in the test result.

1.3.3 Pathways' Significance Test Calculation

A biological pathway is a sequence of interaction among molecules in a cell that leads to a certain product or a change in a cell. The molecules include different genes and miRNAs. Identifying what proteins, mRNA and miRNAs behave abnormally in certain biological pathways can provide us clues about what goes wrong when a disease strikes. For example, given a set of up-regulated and differentially expressed genes under certain diseases, we can perform gene enrichment analysis to find out how many of these genes are enriched in certain pathways and further understand the underlying biological process of the interaction of these genes.

Gene enrichment analysis is a method to identify genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes. The purpose of gene enrichment analysis is to retrieve a functional profile of that gene set with the goal to

better understand the underlying biological processes or molecular functions. Researchers use statistical approaches to identify significantly enriched groups of genes. For our project, we used a public database called Reactome[14,15] to perform gene enrichment analysis to find the significantly enriched pathways. The database calculates the probability of having k significant genes in a pathway via hypergeometric distribution. The p-value of finding *k* significant genes in a pathway is calculated as follows:

$$P = 1 - \sum_{r=0}^{k} \frac{\binom{S}{i} + \binom{N-S}{m-i}}{\binom{N}{m}}$$

N = total number of genes in the dataset

m = number of genes in the pathway

S = number of significant genes in the dataset

k = number of significant genes in the particular pathway

1.3.4 Combine Pathways' P-values using Fisher's Method

Since our goal is to integrate multi-omics data to find disease related pathways. We performed enrichment analysis on diseases specific proteins, mRNAs and also the proteins and mRNA which are significantly and negatively regulated by their target miRNAs to find pathways. Then, we find the common pathways obtained from these 4 platforms which are negatively with their corresponding miRNAs and attempt to integrate the pathways' p-value across the four platforms using Fisher's method.

Fisher's method is a statistical approach which allows us to combine p-values from different independent statistical tests into a single overall test.

$$X^2 = -2\sum_{i=1}^{k} \ln(p_i)$$

p_i is the p-values for the i^{th} hypothesis test. When the p-values tend to be small, the test statistic X^2 will be large, which suggests that the null hypotheses are not true for every test. When all the null hypotheses are true, and the p_i (or their corresponding test statistics) are independent, X^2 has a chi-squared distribution with 2*k* degrees of freedom, where *k* is the number of tests being combined.

The combined pathways' p-value of disease specific mRNA, disease specific proteins and significant mRNA and proteins negatively correlate with their corresponding mRNA and proteins will be applied with Fisher's method to enhance the statistical power. If the q-values of the combined pathways is less than 0.09, we claimed that these pathways are significant. Traditionally, people use 0.05 as q-value threshold. Since our method of using OVE-PUG to obtain disease specific proteins and mRNAs is more stringent, only a few significant genes passed the OVE-PUG test. Due to small number of disease specific genes we found, not that many disease related pathways are found. Hence, we loosen up the q-values' threshold to expand the result of significant disease related pathways we obtained from Fisher's method.

1.4 Experiments

Our dataset consists of 7 groups of different muscular diseases and 1 normal control group with no disease. The 7 muscular diseases type are: DMD, BMD, calpain 3, anoctamine 5, ryanodine receptor, collagen VI, lamin A/C (LMNA). However, we will only be focusing on studying DMD and LMNA for our project since both muscular dystrophies show strong fibrotic signals. We are interested in finding out the molecular pathways that are related to each of the two

diseases using the dataset we have in hand and identify the pathways that are unique to each of the 2 diseases. Our dataset contains 2,171 proteins, 24,355 mRNAs and 870 miRNAs. Each disease and normal group consists of sample sizes of 5 patients. Due to the number of low sample size (n=5) there is for each disease and normal group. We deleted proteins which contain missing values for any of the samples. We also delete mRNAs and miRNAs which has too low of a molecular expression for all 5 samples. If a molecular has an extremely low expression across all 5 samples, we considered the low expression noises produced by the machine. Hence, to avoid false positives, a molecular with extremely low expression is removed from our analysis. After removing protein, mRNAs and miRNAs with missing values and low expression, we are now left with 931 proteins, 870 miRNAs and 20507 miRNAs. A brief summary of how we obtain pathways using the integration of the 3 omics data is shown as flowchart below:

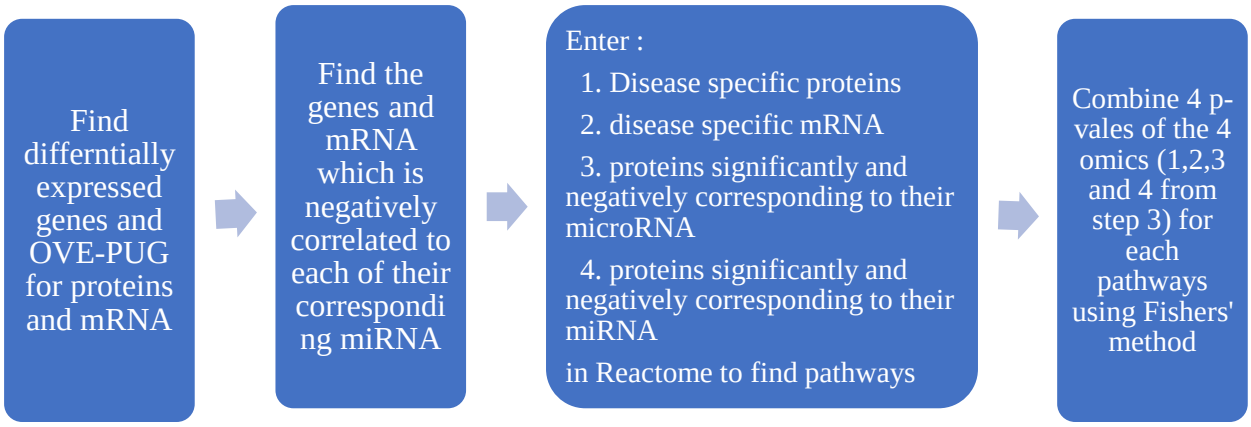

Figure 1. Flowchart of steps

For the first step, we applied three approaches to find disease specific genes. The three approaches are OVE-PUG, T-test and sigmoid function. The result of up-regulated disease specific proteins and mRNA for both LMNA and DMD using the three different methods are shown below. As we can see, we can find the most number of disease specific genes using T-test

and least number of disease specific genes using the OVE-PUG methods. Also, there are more disease specific DMD genes than LMNA genes. We recognize that using the sigmoid function to find differentially expressed genes are not very stringent. Moreover, the threshold and parameter scaling term is not strictly defined. Hence, we will only be using OVE-PUG and T-test method for finding disease related pathways.

Q-values	**DMD**	**norm**	**other**
Q < 0.01	199	1	1
Q < 0.03	313	4	1
Q < 0.05	379	8	2
Q < 0.08	421	14	7
Q < 0.1	450	16	10

Q-values	**LMNA**	**norm**	**other**
Q < 0.01	0	0	0
Q < 0.03	0	0	0
Q < 0.05	0	0	0
Q < 0.08	0	0	0
Q < 0.1	0	0	0

Table 1. OVE-PUG of DMD and LMNA proteins

Q-values	**DMD**	**norm**	**other**
Q < 0.01	0	0	1
Q < 0.03	0	0	2
Q < 0.05	6	0	7
Q < 0.08	370	21	73
Q < 0.1	660	47	115

Q-values	**LMNA**	**norm**	**other**
Q < 0.01	0	0	0
Q < 0.03	2	0	0
Q < 0.05	8	0	6
Q < 0.08	31	3	6
Q < 0.1	36	4	7

Table 2. OVE-PUG of DMD and LMNA mRNA

Q-values	**DMD Vs. Norm**	**DMD Vs. Others**	**Intersect**

Q-values	**LMNA Vs. Norm**	**LMNA Vs. Others**	**Intersect**

Q < 0.01	267	342	208	Q < 0.01	1	12	0
Q < 0.03	434	420	354	Q < 0.03	36	27	4
Q < 0.05	486	445	384	Q < 0.05	178	43	25
Q < 0.08	534	464	418	Q < 0.08	336	69	47
Q < 0.1	554	474	429	Q < 0.1	393	80	57

Table 3. Up-regulated DEG of DMD and LMNA proteins from T-test results

Q-values	**DMD Vs. Norm**	**DMD Vs. Others**	**Intersect**	**Q-values**	**LMNA Vs. Norm**	**LMNA Vs. Others**	**Intersect**
Q < 0.01	1788	2853	789	Q < 0.01	0	103	0
Q < 0.03	6077	4418	2946	Q < 0.03	0	285	0
Q < 0.05	7775	5320	3922	Q < 0.05	3062	437	206
Q < 0.08	9042	6464	4925	Q < 0.08	6221	720	478
Q < 0.1	9571	7206	5393	Q < 0.1	7373	1011	733

Table 4. Up-regulated DEG of DMD and LMNA mRNA from T-test results

Sigmoid scaled parameter	**DMD proteins**	**DMD mRNA**	**LMNA proteins**	**LMNA mRNA**
0.1	389	6157	75	5525
0.3	27	1421	3	779
0.5	3	339	0	167
1	0	19	0	24

Table 5. Up-regulated DEG of DMD and LMNA mRNAs from scaled sigmoid function with threshold > 0.9

Secondly, we calculated the proteins and mRNAs which are significantly and negatively correlated proteins and mRNAs with their corresponding miRNAs. Table 6, 7 and 8 are the numbers of significant proteins and mRNA which are negatively correlated with their corresponding miRNAs we found using target scores from miRDB, correlation coefficient and hypergeometric distribution. As we can tell, we found out a lot more negatively correlated proteins and mRNA using the target score from miRDB compared to the two methods. This is possibly because we did not use any hypothesis test and multiple test correction to find out the genes which are truly targeted by their miRNA targets. Instead, we only considered the target score by the database for solving multiple miRNA-to- one gene problem. We recognized that this approach lacks sensitivity and reliability. Hence, we will not be using the result of this method into find pathways.

DMD proteins	**DMD mRNA**	**LMNA proteins**	**LMNA mRNA**
307	5132	132	1679

Table 6. Proteins and mRNA which are significantly negatively correlated to their corresponding miRNA DEG of DMD and LMNA mRNA using target score from TargetScan

Q-values	**DMD proteins**	**DMD mRNA**	**LMNA proteins**	**LMNA mRNA**
Q < 0.01	1	0	0	0
Q < 0.03	2	0	0	0
Q < 0.05	6	4	0	0
Q < 0.08	11	20	2	0
Q < 0.1	15	44	3	0

Table 7. Proteins and mRNA which are significantly negatively correlated to their corresponding miRNA DEG of DMD and LMNA mRNA correlation coefficient

Q-values	**DMD proteins**	**DMD mRNA**	**LMNA proteins**	**LMNA mRNA**
Q < 0.01	28	787	24	549
Q < 0.03	37	1105	26	757
Q < 0.05	41	1372	30	902
Q < 0.08	39	1647	41	1070
Q < 0.1	55	1827	47	1181

Table 8. Proteins and mRNA which are significantly negatively correlated to their corresponding miRNA DEG of DMD and LMNA mRNA using hypergeometric distribution

In the table below, for DMD and LMNA, we showed the number of pathways found using different approaches of disease specific proteins and mRNA as well as the proteins and mRNA which are negatively correlated to their corresponding miRNA. We also showed the number of significant pathways we obtained after combining the 4 pathways' p-value. The significant pathways here are selected if the combined p-values has q-value less than 0.05 after FDR correction. Since we cannot find any disease specific proteins using OVE-PUG and negatively correlated mRNA with miRNA for LMNA. Hence, we did not submit the LMNA genes we found using these method into Reactome to find pathways.

Disease	Method Finding Disease Specific genes	Method Finding negatively correlated genes - miRNA	Pathways Found in Reactome	Significant Pathways in Found Pathways
DMD	T-test	Negative correlation	32	12
DMD	T-test	Hypergeometric Distribution	215	74
DMD	OVE-PUG	Negative correlation	29	1
DMD	OVE-PUG	Hypergeometric Distribution	200	11
LMNA	T-test	Hypergeometric Distribution	83	23

Table 9. Results of Pathways found in Reactome

We examine each the significant pathways we found in the above 5 scenarios and manually examine which pathways among the significant pathways are related to muscular dystrophy. The results are shown below in Table 10.

Disease	Method Finding Disease Specific Genes	Method Finding negatively correlated genes - miRNA	Name of Pathway Related to Muscular Dystrophy
DMD	T-test	Negative correlation	1. Protein folding 2. Folding of actin by CCT/TriC

			3. Chaperonin-mediated protein folding
DMD	T-test	Hypergeometric Distribution	1. Folding of actin by CCT/TriC 2. Cooperation of Prefoldin and TriC/CCT in actin and tubulin folding 3. Smooth Muscle Contraction 4. Striated Muscle Contraction
DMD	OVE-PUG	Hypergeometric Distribution	RHO GTPase Effectors
LMNA	OVE-PUG	Hypergeometric Distribution	Metabolism of proteins

Table 10. Significant pathways found related to muscular dystrophy

1.5 Discussion

For this project, we demonstrated that it is possible to find disease related pathways using the integration of multi-omics data. Not only were we able to find disease specific genes through OVE-PUG, T-test and sigmoid function, we also showed that multiple methods to solve the problem of solving multiple miRNA regulating one gene problem. OVE-PUG is a more stringent method to find disease specific genes compared to using T-test and sigmoid function. Using hypergeometric distribution to find proteins and mRNA which are negatively correlated also seem to show promising results in the number of proteins and mRNAs we found. This is because our experimental expectation is that DMD is a more severe muscular dystrophy compared to LMNA. Hence, more proteins and mRNAs should negatively regulated by miRNAs

in DMD compared to LMNA. The result of the proteins and mRNAs which we found are negatively correlated with their corresponding miRNAs fit the experimental expectation very nicely. Hence, we can conclude that our method is capable of detecting disease related pathways.

Chapter 2: Personalized Recommendation System for Cancer Treatment Rating Prediction

2.1 Introduction

In the present day, industry is locked in a constant cycle of satisfying consumer needs and turning a profit. One solution that industry has come up with is offering a plethora of options to the consumer so the consumer always has what he needs. Chipotle offers a variety of options that lets the consumer choose exactly what he wants. However, a problem arises with this approach. Too many options leads to choice paralysis. To remedy this, companies like Amazon created recommender systems that personalized the consumer choices from the millions of options they have. This saves the consumer time and also allows industry to customize their inventory to be the most cost efficient because they know what items people will most likely purchase. For health care, there are also too many options for consumers to choose from, and the information is not transparent enough for consumers to make informed choices. A recommender system for cancer treatment is necessary to reduce costs for healthcare, reduce the burden on the consumers

to research the incomprehensible information on medical treatment, and save time for consumers who are only looking for treatment that will work.

For the project, we would like to develop a recommendation system which recommend personalized treatments to cancer patients. Instead of using only the rating information itself, we would also like to incorporate the patient's genetic profile to provide an even better treatment ratings' prediction for patients.

2.2 Background

There are many different existing matrix imputation techniques for filling in missing values in a matrix. For developing recommendation systems, two types of methods are widely used: neighbor-based approach and model-based approach. For neighbor based approach, the main idea is to find similar users[4,5] or items for recommendations[6,7]. More specifically, user-based approaches predict the ratings of active users based on the ratings of similar users found, while item-based approaches predict the ratings of users based on similar items found. The user-based approach recommendation system is the technique Netflix[8] uses to recommend unwatched movies to users. The item-based approach was invented at Amazon[9] to address their large scale challenges with user-based filtering since Amazon has tens of millions of customers and products. Some common techniques of implementing user-based and item-based recommendation systems are Pearson Correlation Coefficient (PCC) and cosine similarity as similarity computation methods.

In contrast of neighbor-based approach, the model-based approach assumes that there is some hidden patterns in the observed user-item ratings. Instead of directly manipulating the original rating database as the neighborhood-based approaches do, model-based approaches build the

latent model based on the observed rating and attempt to use this latent model to uncover the missing rating. Model based approach is preferable than user-based approach for recommendation system because it offers the benefit of both speed and scalability.

One of the most model–based approaches are is low rank matrix factorization(LRMF). The key idea of LRMF is that there exist latent structures in the data. Hence, by uncovering these latent structures, we can potentially recover the missing values in the data.

The common aspect of the approaches discussed above only used the observed ratings for missing rating prediction. While these approaches have been proved successful, using only the rating matrix for prediction only provides the general consensus on the recommendation itself but does not give an insight to the individuals' taste. Hence, we implemented a method which incorporates the user profile into predicting missing rating. By providing additional info that is not available simply from the rating matrix can further personalize a recommendation for users.

2.3 Methodology

In order to see whether the incorporation of patient's genetic data help with treatment response recommendation, we proposed three different methods to impute missing rating prediction. The first and second method are to only use the rating matrix and genetic profile matrix for missing rating imputation. The third method is to incorporate both matrix for rating prediction.

2.3.1 Impute Missing Rating Matrix Using Maximum Margin Matrix Completion

Low rank matrix completion is a commonly used method for imputing missing values in a matrix. One example would be the movie recommender system on Netflix: where the users are recommended to watch certain new movies based on the movies the users watched in the past as

well as other users who watch similar movies as this user. Given a rating matrix $R \in \mathbb{R}^{m \times n}$, m is the numbers of users and n is the number of items. Assuming that R is low rank, we can decompose the R matrix into P and Q matrix, where P is the $m \times k$ user latent matrix, Q is the $n \times k$ item latent matrix and $k \ll \min(m, n)$. The definition of low rank is that there is a latent structure among the items such that the items can be grouped into different categories . That item latent matrix Q suggests that items can be grouped into a small number of different types, and the user latent matrix P represents the affinity of the users to the different movies genres. Our goal here is to adopt maximum-margin matrix factorization method to seek the approximation of R so that

$$R \approx PQ^T$$

where $P \in \mathbb{R}^{m \times k}$ and $Q \in \mathbb{R}^{n \times k}$ with $k < \min(m, n)$.

For a R, let $\Omega \subset \{1,...,m\} \times \{1,...,n\}$ denote the indices of observed entries. The optimization problem we are trying to solve here is

$$\min_{P,Q} \sum_{(i,j) \in \Omega} \left(R_{ij} - (PQ^T)_{ij}\right)^2 + \frac{\lambda}{2} \left(\|P\|_F^2 + \|Q\|_F^2\right)$$

where $\|P\|_F^2$ and $\|Q\|_F^2$ is the Frobenius norm or the sum of squares of entries P and Q. They are also the penalty terms (a.k.a regularization term) to avoid overfitting.

The maximum-margin matrix factorization method is a novel method proposed by Srebro [1]. Traditionally, low rank matrix factorization constraints the rank of the matrix using the nuclear norm of R. However, in maximum-margin matrix factorization, it regularizes the factorization by constraining the norm of U and V. The problem with using low rank factorization is that it can potentially lead to non-convex optimization problems with many local minima. To solve this

problem, low-norm factorization problem uses the L2 constraints which lead to convex optimization problems and can be solved as a semi-definite program.

The loss function above can be solved using alternating least square (ALS), which is a two-step iterative optimization process to solve matrix imputation problem. For ALS algorithm, we take the derivative of loss function with respect to P and Q individually. To find the local minimum of P and Q, we iteratively take the derivative of the loss function with respect to P and Q and solve for P and Q by setting the derivative to 0. In every iteration, we first fix Q and solve for, then fix the newly solved P and solve for Q. In each step the loss function can either decrease or stay unchanged, but never increase. This alternating process will guarantee reduction of the cost function, until convergence.

$$\frac{\partial L}{\partial P_i} = \sum_{j=1}^{n} -I_{ij}(R_{ij} - P_i {Q_j}^T)Q_j + \lambda P_i$$

$$0 = -(R_i - P_i Q^T)Q + \lambda P_i$$

$$P_i(Q^T Q + \lambda I) = R_i Q$$

$$P_i = R_i Q(Q^T Q + \lambda I)^{-1}$$

The derivation of Q is calculated in the similar manner as P.

$$\frac{\partial L}{\partial Q_j} = \sum_{i=1}^{m} -I_{ij}(R_{ij} - P_i {Q_j}^T){P_i}^T + \lambda Q_j$$

$$0 = -\left(R_j - P {Q_j}^T\right)P^T + \lambda Q_j$$

$$Q_j(P^T P + \lambda I) = R_i P^T$$

$$Q_j = R_i P^T (P^T P + \lambda I)^{-1}$$

Similar to gradient descent optimization, ALS is guaranteed to converge only to a local minimum, and it ultimately depends on the initial values for P or Q. However, if the problem fortunately is convex, the local minimum will also be the global minimum.

2.3.2 Impute Missing Rating Using Only Patients' Similarity from Genetic Profile

Since the rating matrix is relatively sparse, we want to see if patients' genetic profile alone can be used for rating prediction. We computed the nearest neighbors of each patient based on the similarity between patients in the genetic profile. To compute the missing treatment rating for a specific patient, we take the average of the treatment rating of this patients' top nearest neighbors who receive the treatment before as its predicted rating. For example, we are interested in predicting the missing treatment *R1* for patient *P1*. We would take the average ratings of *P1*'s top 5 nearest neighbors who has receive the treatment *R1* as *R1*'s prediction value for *P1*. The assumption here is that users with similar genetic profile will respond similarly to the same time type of treatment. Hence, by finding out the ratings of the neighbors of the users with missing treatment ratings, we can further predict the missing rating of each user. The neighbor of a user is defined by the similarity between one user and other users using their genetic profile. The similarity is calculated using cosine similarity and Jaccard similarity. We tried two different kinds of similarity matrix to see which one of them help increase the prediction accuracy of the missing rating more.

The first similarity matrix is calculated using cosine similarity.

$$Cosine\ Similarity(i, f) = \frac{\sum_{k=1}^{n} A_k B_k}{\sqrt{\sum_{k=1}^{n} A_k} \sqrt{\sum_{k=1}^{n} B_k}}$$

where A_k and B_k are the gene expression k for patient i and n is the total number of patients in the dataset.

The second similarity is calculated using Jaccard similarity.

$$Jaccard\ Similarity(i, f) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

where A and B are the gene expression for patient i.

2.3.3 Combine Patients' Rating and Genetic Information Matrix

Previously, we proposed two different approaches to solve the missing value imputation problem by finding similar patterns in either treatments rating or patient's genetic profile. However, these two approaches are insensitive to treatment rating prediction for those patients who have similar treatments response but different genetic profile. To overcome this problem, we adopt a method which uses another similarity regularization term to impose constraints between each patient and the rest of other patients individually. This method is originally proposed by Dr. Hao Ma from Hong Kong University to improve matrix factorization algorithms by incorporating social network information for social-based recommender systems. They proposed several methods to model social network information as regularization terms to constrain the matrix factorization framework. One of the problem they aimed to tackle was the inaccurate modeling of feature vector P_i due to users with friends who have diverse taste. Hence, they proposed to add another social regularization term to the low norm factorization algorithm in Section 1.3.1 to impose constraints between one user and their friends individually.

$$\frac{\beta}{2}\sum_{i=1}^{m}\sum_{f=1}^{m} Sim(i, f)\left\|P_i - P_f\right\|_F^2$$

$$Sim(i,f) = \frac{\sum_{j \in I(i) \cap (f)} R_{ij} \cdot R_{fj}}{\sqrt{\sum_{j \in I(i) \cap (f)} {R_{ij}}^2} * \sqrt{\sum_{j \in I(i) \cap (f)} {R_{fj}}^2}}$$

Where $\beta > 0$, and $Sim(i,f)$ is the same similarity function computer using Person's correlation coefficient on item rated by both users. Item j belongs to the subset of items which user i and user f both rated. R_{ij} is the rate user i gave item j.

Since we are more interested in seeing how the patients' similarity in genetic expression would impact treatment rating prediction, we compute the similarity function on patient's genetic profile G. Moreover, since our genetic profile is in binary form, cosine similar seems to be a more appropriate measure for our similarity function. The similarity function for user i and f to cosine similarity using patients' genetic profile is shown below:

$$Sim(i,f) = \frac{\sum_{k=1}^{h} G_{ik} G_{fk}}{\sqrt{\sum_{k=1}^{h} G_{ik}} \sqrt{\sum_{k=1}^{h} G_{fk}}}$$

Where h is the total amount of genes in the genetic profile, G_{ik} is the genetic expression for user i on gene k and G_{fk} is the genetic expression for user f on gene k.

As we can see, a small value of $Sim(i,f)$ indicates that the distance between user latent vectors P_i and P_f should be larger, while a large value tells that the distance between the feature vectors should be smaller.

Now, our modified matrix imputation problem can be remodeled as:

$$\min_{P,Q} L(R,P,Q) \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij} \left(R_{ij} - {P_i}^T Q_j\right)^2 + \frac{\beta}{2} \sum_{i=1}^{m} \sum_{f=1}^{m} Sim(i,f) \left\| P_i - P_f \right\|_F^2$$

$$+ \frac{\lambda}{2} \left(\|P\|_F^2 + \|Q\|_F^2 \right)$$

Where λ and β are the tuning parameters for the regularization term.

Another advantage of this approach is that the added regularization term in the loss function above provides a better estimate of the distance between users. More specifically, if user i has similar genetic expression and treatment response as user f, the distance between latent vector P_i and P_f would be small. On the other hand, if user i has very different genetic expression yet similar treatment response as user f, the distance between latent vector P_i and P_f would be larger due to a larger distance in genetic profile and smaller distance in treatment responses are balanced out for user i and f.

Similar to our first approach of using only rating matrix to impute missing rating, a local minimum of the objective function above can also be found by performing gradient descent or alternating least square algorithm in latent feature vectors P_i and Q_i. The derivation of the loss function is shown below.

$$\frac{\partial L}{\partial P_i} = \sum_{j=1}^{n} -I_{ij}(R_{ij} - P_i Q_j)Q_j + \lambda_p P_i + \beta \sum_{f=1}^{m} Sim(i,f)(P_i - P_f)$$

$$\frac{\partial L}{\partial Q_j} = \sum_{i=1}^{m} -I_{ij}(R_{ij} - P_i Q_j)P_i + \lambda_Q Q_j$$

$$P_i' = P - \alpha \frac{\partial L}{\partial P_i}$$

$$Q_j' = Q - \alpha \frac{\partial L}{\partial Q_j}$$

where α is the rate of approaching the minimum.

2.4 Experiments

Both treatment response and genetic profile datasets are provided by Perthera, Inc. The rating in the treatment response dataset represents the progression-free survival(PFS) for 5 different drugs. The rating ranges from 1 to 4. Each value represents the length of the time the disease system worsen after a patients receive the treatment for the disease. Moreover, the genetic profile is the gene mutation data expressed in binary form, where 1 represents an existence of mutation and 0 represents no found mutation. To evaluate the purpose of each of the matrix imputation techniques, we use 5-fold cross validation method to calculate the root mean square error(RMSE) of the predicted ratings and their reference rating. The RMSE rating for each of the algorithm is shown below. As we can see, the RMSE of integrating patients' genetic profile for rating prediction is slightly lower than the other 2 algorithms. Hence, we can conclude that integrating patients' genetic profile indeed help with the prediction accuracy. However, since the RMSE diseases as the rank of the matrix increases, it is possible that rating matrix is not truly low rank.

Rank	**RMSE for treatment rating only**	**RMSE for treatment rating and genetic profile**
2	1.06	0.92
3	0.984	0.90
4	0.87	0.85
5	0.88	0.83

Table 11. RMSE for using treatment rating only and with genetic profile combined

neighbors	RMSE for genetic profile only using Jaccard similarity	RMSE for genetic profile only using cosine similarity
5	1.41	1.21
10	1.44	1.22
15	1.46	1.22

Table 12. RMSE for using genetic profile only

2.5 Discussion

For this project, we demonstrated that incorporating patient's genetic profile can improve the prediction accuracy compared to using the rating matrix alone for imputation. One possible reason why the RMSE decreases as the rank increase is possibly due to the fact that the missing ratings are not random. This happens because our treatments can be grouped into first and second line of treatment. First line therapy is the treatment given to patient typically at the early of a disease. The intent of first-line therapy is to cure the disease as early as possible before the disease becomes worse. A second line treatment is given to patients to improve the poor outcome of the first line treatment. Most patients in our dataset receive treatments in sequential order. Most patients who receive first-line treatment have not yet receive the second-line treatment. Hence, we have more patients who receive rating of the first line but not the second line treatment. For our next step, we would like to solve this problem and hopefully see a increase in prediction accuracy.

Chapter 3: Contribution of Work

For the thesis, we introduce novel approaches to integrate multi-omics data (miRNA, proteins and mRNAs) to detect disease related pathways. The approaches include a 4-step based statistical approach to accurately identify disease related pathways. Moreover, we also develop a personalized recommendation system to impute missing treatment response with incorporating patient's genetic profile for cancer patients.

References

[1] Godard, Patrice and Jonathan van Eyll. "Pathway Analysis from Lists of Micrornas: Common Pitfalls and Alternative Strategy." *Nucleic Acids Research*, vol. 43, no. 7, 2015, pp. 3490-3497, PMC, doi:10.1093/nar/gkv249.

[2] Hung, Jui-Hung et al. "Identification of Functional Modules That Correlate with Phenotypic Difference: The Influence of Network Topology." *Genome Biology*, vol. 11, no. 2, 2010, p. R23, doi:10.1186/gb-2010-11-2-r23.

[3] Seo, Jiyoun et al. "Integration of Microrna, Mrna, and Protein Expression Data for the Identification of Cancer-Related Micrornas." *PLOS ONE*, vol. 12, no. 1, 2017, p. e0168412, doi:10.1371/journal.pone.0168412.

[4] Redpath, Jennifer et al. "User-Based Collaborative Filtering: Sparsity and Performance." *Proceedings of the 2010 conference on STAIRS 2010: Proceedings of the Fifth Starting AI Researchers' Symposium*, IOS Press, 2010, pp. 264-276.

[5] Breese, John S. et al. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering." *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 1998, pp. 43-52.

[6] Deshpande, Mukund and George Karypis. "Item-Based Top-<I>N</I> Recommendation Algorithms." *ACM Trans. Inf. Syst.*, vol. 22, no. 1, 2004, pp. 143-177, doi:10.1145/963770.963776.

[7] Sarwar, Badrul et al. "Item-Based Collaborative Filtering Recommendation Algorithms." *Proceedings of the 10th international conference on World Wide Web*, ACM, 2001, pp. 285-295. doi:10.1145/371920.372071.

[8] Gomez-Uribe, Carlos A. and Neil Hunt. "The Netflix Recommender System: Algorithms, Business Value, and Innovation." *ACM Trans. Manage. Inf. Syst.*, vol. 6, no. 4, 2015, pp. 1-19, doi:10.1145/2843948.

[9] Linden, G. et al. "Amazon.Com Recommendations: Item-to-Item Collaborative Filtering." *IEEE Internet Computing*, vol. 7, no. 1, 2003, pp. 76-80, doi:10.1109/MIC.2003.1167344.

[10] Romero-Cordoba, Sandra L. et al. "Mirna Biogenesis: Biological Impact in the Development of Cancer." *Cancer Biology & Therapy*, vol. 15, no. 11, 2014, pp. 1444-1455, PMC, doi:10.4161/15384047.2014.955442.

[11] Wong, Nathan and Xiaowei Wang. "Mirdb: An Online Resource for Microrna Target Prediction and Functional Annotations." *Nucleic Acids Research*, vol. 43, no. Database issue, 2015, pp. D146-D152, PMC, doi:10.1093/nar/gku1104.

[12] Wang, Xiaowei. "Improving Microrna Target Prediction by Modeling with Unambiguously Identified Microrna-Target Pairs from Clip-Ligation Studies." *Bioinformatics*, vol. 32, no. 9, 2016, pp. 1316-1322, doi:10.1093/bioinformatics/btw002.

[13] Fabregat, Antonio et al. "The Reactome Pathway Knowledgebase." *Nucleic Acids Research*, vol. 44, no. Database issue, 2016, pp. D481-D487, PMC, doi:10.1093/nar/gkv1351.

[14] Croft, David et al. "The Reactome Pathway Knowledgebase." *Nucleic Acids Research*, vol. 42, no. Database issue, 2014, pp. D472-D477, PMC, doi:10.1093/nar/gkt1102.

[15] Yu, Guoqiang et al. "Pugsvm: A Cabigtm Analytical Tool for Multiclass Gene Selection and Predictive Classification." *Bioinformatics*, vol. 27, no. 5, 2011, pp. 736-738, doi:10.1093/bioinformatics/btq721.

[16] Xiao, Yun et al. "Identifying Dysfunctional Mirna-Mrna Regulatory Modules by Inverse Activation, Cofunction, and High Interconnection of Target Genes: A Case Study of Glioblastoma." *Neuro-Oncology*, vol. 15, no. 7, 2013, pp. 818-828, PMC, doi:10.1093/neuonc/not018.